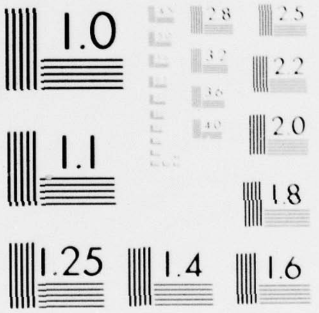
AD-A077 890 NORTH CAROLINA UNIV AT CHAPEL HILL INST OF STATISTICS F/G 12/1
ROBUST METHODS FOR FACTORIAL EXPERIMENTS WITH OUTLIERS.(U)
JUL 79 R J CARROLL AFOSR-75-2796
UNCLASSIFIED MIMEO SER-1244 AFOSR-TR-79-1097 NL

| OF |

AD
A077 890

END
DATE
FILMED
1-80
DDC


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY OF NORTH CAROLINA

1473

Robust Methods for Factorial Experiments with Outliers

by

Raymond J. Carroll
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series #1244

July 1979

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

ROBUST METHODS FOR FACTORIAL EXPERIMENTS WITH OUTLIERS

Raymond J. Carroll*

University of North Carolina at Chapel Hill

Abstract

Two factorial experiments with possible outliers (John (1978)) are reanalyzed by means of robust regression techniques. We show that using M-estimates of regression results in efficient analyses which are easier to implement than the methods proposed by John.

Key Words: Factorial experiment, outliers, robustness, M-estimates

*This research was supported by the Air Force Office of Scientific Research under Contract AFOSR-75-2796.

1. Introduction

In a recent paper, John (1978) discussed the effects and detection of outliers in factorial experiments. His methods were based on least squares methodology and can be quickly summarized as follows. Consider the usual linear model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \quad .$$

If one suspects that m outliers are present, then the model could be written as

$$\underline{Y} = X\underline{\beta} + \sum_{i=1}^{m} \theta_i \underline{d}_i + \varepsilon \quad ,$$

where $\underline{d}_i$ is a vector with 1 in the row corresponding to the i-th suspected outlier and 0 elsewhere (i = 1,2,...,m). The presence of outliers would then be tested by H_0: $\underline{\theta} = 0$. However, the percentage point F_α which one would compare with the F-statistic F* should not be the usual upper α percentage point of the F distribution since the observations being tested have the most extreme residuals.

John then discusses a simple modification when there is only one suspected outlier (m = 1 — use the upper α/N percentage point of the F-distribution). For two suspected outliers (m = 2), a fairly deep analysis is required, including the use of simulation. For $m \geq 3$, the method becomes even more detailed.

At best, the type of analysis presented above is messy, time-consuming and complex. The method of determining the percentage point F_α is neither unified nor written as a simple algorithm, and outlier rejection itself requires skill and care. Practicing statisticians with deadlines to meet will likely not have the time, inclination nor energy to do such an analysis, while the common naive users of canned statistical programs (such as SPSS) will not have the technical competance to do the analysis, if indeed he or she even bothers to look for outliers.

The basic difficulty is that least squares is sensitive to outliers, which can drastically change parameter estimates as well as disguise significant effects; this is amply illustrated by John's first example (see Section 3 below). In order to improve the quality of the mass of statistical analyses, what is needed are not complex, *ad hoc* procedures for detecting and modifying outliers, but rather procedures which are insensitive to the presence of outliers (see Cook (1977) for a treatment of outliers in independent variables in regression). Ideally, such robust (non-least squares) methods should give results (for most problems!) which are similar to those obtained by outlier detection and modification procedures. The recent statistical literature abounds with proposals to attain this goal, the most popular of which are Huber's (1964, 1973, 1977) M-estimates. This rather intensively studied class has been specifically designed to be insensitive to outliers and to retain high efficiency when the errors are heavier-tailed than the normal, two properties not possessed by least squares. The basics of M-estimates are reviewed in Section 2. In Section 3 we apply these methods to John's first example, while in Section 4 we discuss John's second example. We find in both cases that one pass through a robust regression program based on M-estimates yields results closely similar to those obtained by John's complex analysis.

Daniel and Wood (1971) and Andrews (1974) analyze a data set in a regression context. The former use least squares and, decide (after very detailed analysis) that there are three outliers. Andrews uses robust techniques similar to those presented here and shows that the decision favoring three outliers can be reached in a much easier and more routine fashion. Thus, the advantages of robust techniques are not limited to the factorial experiments analyzed herein.

2. M-estimates of Regression

Least squares estimates minimize

(2.1) $$\sum_{i=1}^{n} \rho((y_i - \underline{x}_i\beta)/\sigma) \quad ,$$

where $\rho(x) = \frac{1}{2}x^2$. The quadratic form of ρ is what makes least squares sensitive to outliers. This can also be seen if one defines $\psi = \rho'$, for then one solves

(2.2) $$\sum_{i=1}^{n} \psi((y_i - \underline{x}_i\beta)/\sigma)\underline{x}_i = 0 \quad ,$$

where for least squares $\psi(Z) = Z$. In order to achieve robustness against outliers and high efficiency for distributions heavier-tailed than the normal, Huber (1964), Andrews, et al (1972) and Hampel (1974) suggest that ψ be a bounded function, and that scale be estimated in one of two ways:

(Proposal 2) Solve simultaneously (2.2) and

(2.3) $$(n-p)^{-1}\sum\psi^2((y_i - x_i\beta)/\sigma) = E\psi^2_\Phi(Z) \quad ,$$

the expectation being under the standard normal,

(2.4) $$(\text{MAD})\hat{\sigma} = \left| \begin{matrix} \text{median absolute residual} \\ \text{from median} \end{matrix} \right| /.6745$$

(This is asymptotically equal to one for the normal model.)

In both cases, the solution is found iteratively. One chooses a starting value for σ, solves (2.2), then updates σ by (2.3) or (2.4), etc., continuing until convergence. Algorithms are available in Huber (1973, 1977) and Dutter (1976); the author has adapted these algorithms for use in the SAS computer programs (a card deck is available upon request). In neither case is the computation burdensome.

The typical choices of ψ are

Huber's $\psi(x) = -\psi(-x) = \max(-k, \min(x,k))$, with k generally taken as 1.5 or 2.0.

Hampel's
$$\begin{aligned} \psi(x) &= -\psi(-x) & \\ &= x & 0 \le x < a \\ &= a & a \le x < b \\ &= a\frac{(c-x)}{(c-b)} & b \le x \le c \\ &= 0 & x > c\ . \end{aligned}$$

Andrews'
$$\begin{aligned} \psi(x) &= -\psi(-x) & \\ &= \text{sine}(x/c) & 0 \le x \le c\pi \\ &= 0 & x \ge c\pi\ . \end{aligned}$$

The constant c is often taken as 2.1.

The fact that Hampel's ψ and Andrews' ψ both redescend to zero suggests (Hampel (1974)) that they give no weight to gross outliers, while Huber's ψ will give some weight to these outliers but not nearly so much as least squares. The redescending ψ functions (unlike Huber's ψ) can have problems with convergence; for this reason we adopt the convention of first estimating $\underline{\beta}$ by Huber's method and then using at most two iterations of the algorithm for Hampel's and Andrews' methods. In all these cases, under proper conditions, the robust regression estimate $\hat{\beta}_R$ of $\underline{\beta}$ is asymptotically normally distributed with mean $\underline{\beta}$ and covariance matrix which can be estimated by

$$(2.4) \qquad (n\hat{\sigma}^2\textstyle\sum\psi^2(r_i)/\{\sum\psi'(r_i)\}^2)(X'X)^{-1}\ ,$$

where the standardized residuals are

$$r_i = (y_i - \underline{x}_i\hat{\underline{\beta}})/\hat{\sigma}\ .$$

Huber (1973) and Andrews, et al (1972) show in simulation experiments that the estimates $\hat{\beta}_R$ are generally more efficient than the least squares estimates; they are only slightly more variable than least squares for the normal model but are considerably less variable for heavier-tailed models.

Inference about the parameters can take at least two forms, both based on the approximation (2.4). Schrader and Hettmansperger (unpublished) suggest an analysis of the usual drop in sum of squares statistic using $\sum\rho(r_i)$. Bickel (1976, discussion section) suggests the approach used here. Let

$$\lambda = n^{-1}\sum_{i=1}^{n}\psi'(r_i)$$

$$\eta = 1 + (p/n)(1-\lambda)/\lambda \ .$$

The term η is suggested by Huber (1973, equation) as a variance inflation factor for (2.4) if p/n is not small. Define *pseudo-values*

$$\tilde{Y}_i = \underline{x}_i\hat{\beta}_R + \eta\hat{\sigma}\psi(r_i)/\lambda \ .$$

Then, the least squares estimates for the model $\underline{\tilde{Y}} = X\underline{\beta} + \underline{\varepsilon}$ are exactly $\hat{\beta}_R$ (this follows from (2.2)). Bickel suggests that asymptotically correct tests can be obtained by defining the pseudo-values and using them in conventional least squares packages.

In the examples below, we used $\hat{\sigma}$ given by (2.3); for Huber's ψ we took k = 1.5, for Hampel's a = 1.5, b = 3.5, c = 8.0, while for Andrews', c = 2.1.

3. First Example

John's first example is a 3^{4-1} fractional replication of a 3^4 experiment. The effects of each factor are split into linear and quadratic components (AL, AQ, BL, BQ, CL, CQ, DL, DQ) and three interactions are formed by multiplication (ALBL, ALCL, BLCL). Observation 11 is a suspected outlier; the predicted values and residuals for the four methods are given in Table 1. In Figure 1 we present schematic plots of the residuals for the four methods (see Tukey (1972)). The length of the box corresponds to the interquartile range, the length of the tails is described by the verticle dashed lines, and potentially serious outliers are indicated by the symbol '*'. The obvious conclusion from Table 1 is that the Hampel and Andrews method are particularly robust in that their predicted value for observation 11

(i) Is close to John's refitted value;

(ii) hardly changes when the original observation y - 14 is replaced by the refitted value y = 62.33.

From Figure 1 we see that the robust methods fit the data much better than does least squares, again indicating the value of these procedures.

In Table 2 we present significance levels for the effects using the original observations and then using the refitted observation 11. The striking features are that

(i) The Hampel and Andrews methods give in one pass on the original observations essentially the same analysis as does John's more complex refitted data;

(ii) the significance levels of the Hampel and Andrews methods do not change to any large extent after observation 11 is modified.

We conclude that for the 3^{4-1} example, the robust regression methods compare favorably with John's method. The significance levels are similar and the robust methods are, in general, considerably easier to implement.

4. Second Example

The second example John uses to illustrate his method is a confounded 2^5 experiment, the block effects confounding the highest order interaction. After lengthy analysis he concludes that two suspected outliers are not really outliers and should not be refitted. In Table 3 we present significance levels for the four tests, while in Figure 2 the schematic plot of the residuals is given.

The major difference between least squares and the robust estimates exhibited in Table 3 is that the latter show a main effect in B significant at the .05 level, while for least squares the significance level is approximately .13. Although John's analysis suggests that there may well be no *gross* outliers, we see that the treatment combinations ad and d are sufficiently discrepant from the others so as to inflate the least squares mean square error and thus obscure what appears to be a significant main effect. The Hampel and Andrews methods also find the BE interaction to be (moderately) significant.

In the previous section we found that a gross outlier can radically affect a least squares analysis, while having a much smaller effect on the robust methods. In this example we have seen that slightly discrepant observations (perhaps due to a distribution heavier-tailed than the normal) can inflate the least squares mean square error, causing it to lose efficiency.

5. Discussion

The examples made clear that there is much to be gained by using M-estimates of regression. The treatment of designs (with possible outliers) by these methods is efficient and easy to implement. The quality of statistical analyses will be greatly improved by routine use of M-estimates as one of the statistician's tools.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances.*

Andrews, D.F. (1974). A robust method for multiple linear regression. *Technometrics, 16*, 523-531.

Bickel, P.J. (1976). Another look at robustness. *Scand. J. Statist. 3*, 145-168.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics, 19*, 15-18.

Daniel, C. and Wood, F.S. (1971). *Fitting Equations to Data.* Wiley: New York.

Dutter, R. (1976). LINWDR: Computer linear robust curve fitting program, Res. Rep. #10, Fachgruppe für Stat., Eidenössische Technische Hochschule, Zurich.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc. 69*, 383-393.

Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* *35*, 73-101.

Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* *1*, 799-821.

John, J.A. (1978). Outliers in factorial experiments. *Appl. Statist.* *27*, 111-119.

Tukey, J.W. (1972). Some graphic and semi graphic displays. In *Statistical Papers in Honor of George W. Snedecor*, T.A. Bancroft, ed.

Table 1

Predicted values and residuals for observation 11 in the 3^{4-1} experiment, based on the original observations. John's refitted value is 62.33.

	Least Squares	Huber	Hampel	Andrews
Observed value	14	14	14	14
Predicted value	46.2	55.7	57.7	59.4
Residual	-32.2	-41.7	-43.7	-45.4
Refitted value	62.33	62.33	62.33	62.33
Predicted value after refitting	62.33	61.2	61.2	61.6
Residual after refitting	0	1.1	1.1	0.7

Table 2

Significance levels for the 3^{4-1} experiment using the four methods.

Blank values indicate a significance level greater than 0.10.

	Original Observations (Y_{11} = 14)				Modified Observations (Y_{11} = 62.33)			
Effects	Methods				Methods			
	Least Squares	Huber	Hampel	Andrews	Least Squares	Huber	Hampel	Andrews
AL	.04	.00	.00	.00	.01	.00	.00	.00
AQ		.02	.01	.01	.01	.00	.01	.01
BL			.10	.08	.03	.06	.07	.07
BQ			.06	.09	.07	.10	.10	(.11)
CL								
CQ			.05	.06	.03	.02	.03	.04
DL								
DQ								
ALBL								
ALCL								
BLCL		.00	.00	.01	.01	.00	.00	.01
Predicted Value	46.22	55.70	57.72	59.35	62.33	61.20	61.17	61.60

Table 3

Significance levels for the confounded 2^5 experiment. Blanks indicate a level greater than 0.10.

Source	Least Squares	Huber	Hampel	Andrews
A	.01	.00	.00	.00
B		.03	.02	.02
C	.03	.01	.01	.03
D	.00	.00	.00	.00
E			(.11)	.08
Blocks				
AB				
AC	.04	.04	.04	.07
AD				
AE	.04	.00	.00	.00
BC				
BD				
BE			.09	.06
CE	(.11)	.06	.07	.10
DE				

Figure 1

Schematic plots for the residuals in the 3^{4-1} experiment using the original observations.

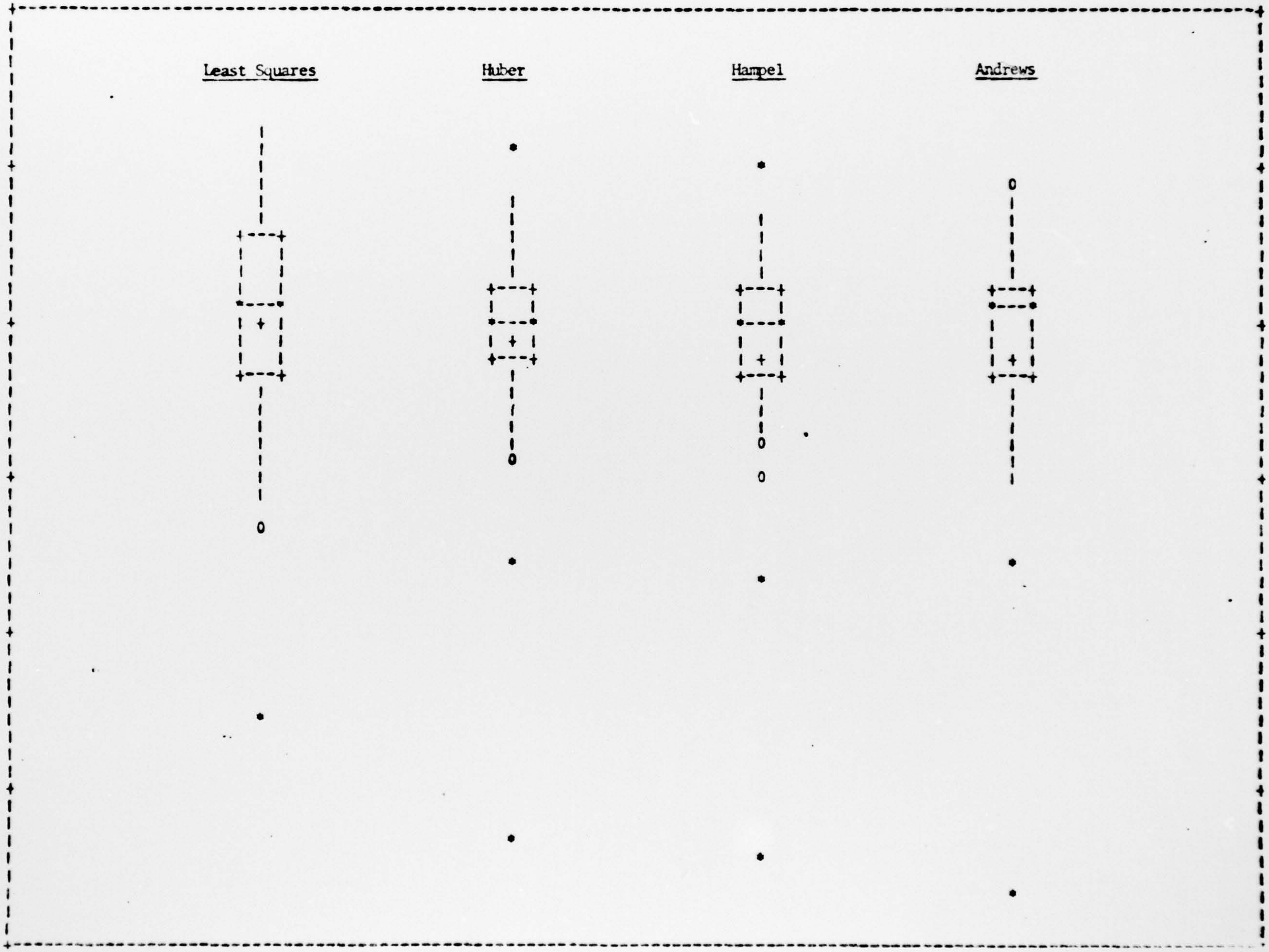
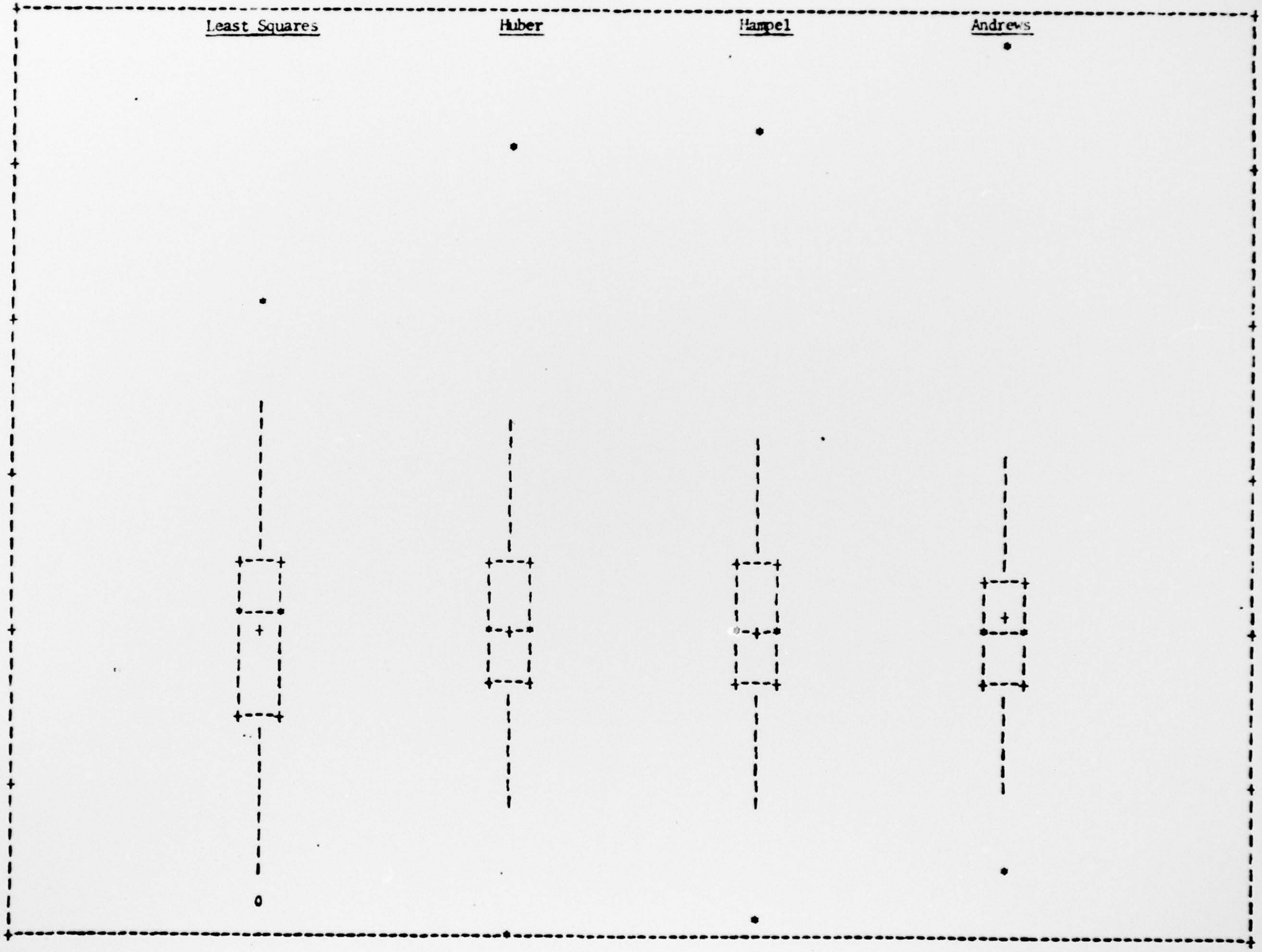


Figure 2

Schematic plots of residuals for the four methods in the confounded 2^5 experiment.



Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (18) (19) AFOSR-TR-79-1097 | 2. GOVT ACCESSION NO. (9) Interim rept.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) Robust Methods for Factorial Experiments with Outliers	5. TYPE OF REPORT & PERIOD COVERED Interim
	6. PERFORMING ORG. REPORT NUMBER (14) Mimeo Series No. 1244
7. AUTHOR(s) (10) Raymond J. Carroll	8. CONTRACT OR GRANT NUMBER(s) (15) ✓AFOSR-75-2796
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of North Carolina Chapel Hill, North Carolina 27514	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (16) (17) 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research /NM (11) Bolling AFB, DC 20332	12. REPORT DATE July 1979
	13. NUMBER OF PAGES 14
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) (12) 16	15. SECURITY CLASS. (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release: Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Factorial experiment, outliers, robustness, M-estimates

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Two factorial experiments with possible outliers ~~(John (1978))~~ are reanalyzed by means of robust regression techniques. We show that using M-estimates of regression results in efficient analyses which are easier to implement than the previous methods proposed by John.

410 064

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)